

Weekly Report

Lu Junhua

2017 年 5 月 15 日

Done

- Learn data structures, complete a dozen of leetcode problems and become more familiar with JavaScript syntax.
- Some discussions on radar projection.
- Design the data attributes used for case study of BeXplorer.

To do

- Test on potential projection methods.
- A preliminary version of 专著.

Papers

- KDD 16 *Large-Scale Item Categorization in e-Commerce Using Multiple Recurrent Neural Networks* 做的是用RNN进行商品分类, 利用的属性不多, 只有商品名称、品牌名、粗分类、用户id、制造商、图像签名(提取的图像特征), 可以说没有太多的先验知识. 这些属性有单词序列也有次序值. 这是一种end to end的方法, 通俗的理解就是给他输入最后会有输出, 中间一些过程诸如模型选择之类就不用操心了. 它对每个属性都单独有个RNN, 结果也有了较大的提高, 我们在以后无法分类时也可以使用. 在评估时, 他们将数据集分为训练集、验证集、和测试集三部分, 这与我们平时的不太一样, 有个经典示例是(训: 分类器权重、验: 分类器架构、测: 评估诸如普适性等指标).
- KDD 16 *Lossless Separation of Web Pages into Layout Code and Data* 文章投在了KDD16, 但是却是做(静态)网页的代码与数据(数据是网页DOM树中的一些value) 分离, 看起来并不是一个特别难的问题. 一方面之前的工作做得没有那么细致、一方面从数据挖掘角度抽象出了很多的准则(或者说损失函数之类, 采用了独特的距离度量如MDL), 对于大规模的页面可能确实非常高效. 具体就是实现代码的压缩表达, 将相同的元素表达抽取出来, 简化了表达; 并且可以支持条件语句的表示(比如某一段压缩的代码仅在某条件下会触发).
- FCS 2016 *Event analysis in social multimedia: a survey* 社交多媒体与事件之间有着复杂的联系, 事件的爆发往往会有社交多媒体做预兆、而多媒体在事件中又会有不同的表现. 综述就是介绍如何发现关联、强化冠词, 从事件增强描述(找关联, 丰富事件信息)、事件检测和事件分类的角度, 提供了大量方法, 与我们平时在可视化界看到的一些文章思路并不相同, 不过也许可以作为后面工作的一些借鉴.

- Machine Learning 03 *Tree Induction for Probability-Based Ranking* 谢聪师兄在VAET中用的决策树, 其实是概率估计决策树(PET)(一个叶节点若有100个点, 90个为正, 那么就说概率为90%), 文章提出的基于概率排序的决策树, 其实在实际生活中很有用. 本文作者之一其实也是我们之前在用的在线决策树的作者之一, 可以说在决策树方面非常有研究. 文章讲的内容非常丰富, 包括由来、不足、不足的根本原因(一般决策树与概率决策树之区别等)、引入集成学习、实际方法评估. 相对于C4.5, 这种树对于概率用了laplacian correction, 并且在生成中对于某种特定的情形不剪枝, 效果很好.